Inference for categorical data

Hsuan-Hao Fan

Feb. 7, 2019

In August of 2012, news outlets ranging from the Washington Post to the Huffington Post ran a story about the rise of atheism in America. The source for the story was a poll that asked people, "Irrespective of whether you attend a place of worship or not, would you say you are a religious person, not a religious person or a convinced atheist?" This type of question, which asks people to classify themselves in one way or another, is common in polling and generates categorical data. In this study, we take a look at the atheism survey and explore what's at play when making inference about population proportions using categorical data.

Getting Started

Load packages

First, we explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. The data can be found in the companion package for this course, `statsr`.

Let's load the packages.

```
library(statsr)
library(dplyr)
library(ggplot2)
```

The survey

The press release for the poll, conducted by WIN-Gallup International, can be accessed here.

A poll conducted by WIN-Gallup International surveyed 51,927 people from 57 countries by using the following methods to gather information: face to face, telephone, and internet. In the first paragraph, several key findings are reported. These percentages appear to be **sample statistics**.

The data

By close looking at Table 6 (pages 15 and 16), which reports the sample size and response percentages for all 57 countries, while this is a useful format to summarize the data, we will base our analysis on the original data set of individual responses to the survey. Load this data set into R with the following command.

```
data(atheism)
```

Note that each row of Table 6 corresponds to **countries**. In contrast, each row of `atheism` corresponds to **individual persons**.

To investigate the link between these two ways of organizing this data, take a look at the estimated proportion of atheists in the United States. Towards the bottom of Table 6, we see that this is 5%. We should be able to come to the same number using the `atheism` data.

First, we create a new dataframe called `us12` that contains only the rows in `atheism` associated with respondents to the 2012 survey from the United States:

```
us12 <- atheism %>%
  filter(nationality == "United States" , atheism$year == "2012")
```

Next, we calculate the proportion of atheist responses in the United States in 2012, i.e. in `us12`.

```
# type your code for Question 7 here, and Knit
us12 %>%
     group_by(us12$response) %>%
     summarize(counts = n())
```

```
## # A tibble: 2 x 2
##   `us12$response` counts
##   <fct>            <int>
## 1 atheist             50
## 2 non-atheist        952
```

Hence, the percentage of atheist responses in the United States in 2012 is 0.0499 ≈ 5%. This percentage agrees with the percentage in Table 6.

Inference on proportions

As was hinted earlier, Table 6 provides **sample statistics**, that is, calculations made from the sample of 51,927 people. What we'd like, though, is insight into the population **proportion parameters**. We'd like to answer the question, "What proportion of people in our sample reported being atheists?" with a statistic; while the question "What proportion of people on earth would report being atheists" is answered with an estimate of the parameter.

We can use the inferential tools for estimating population proportion to calculate the confidence interval and the hypothesis test.

Calculate confidence intervals for the proportion of atheists in 2012 in United States

1. Check conditions for inference

First of all, we check the conditions for inference to construct a 95% confidence interval for the proportion of atheists in United States in 2012.

- Since the study uses stratified sampling, and the sample size is 1,002 less than 10% of popuation, it is reasonable to assume that observations are **independent**.
- There are 50 successes (atheist) and 952 failures (non-atheist). They are both larger than 10 so **the success-failure condition is satisfied.**

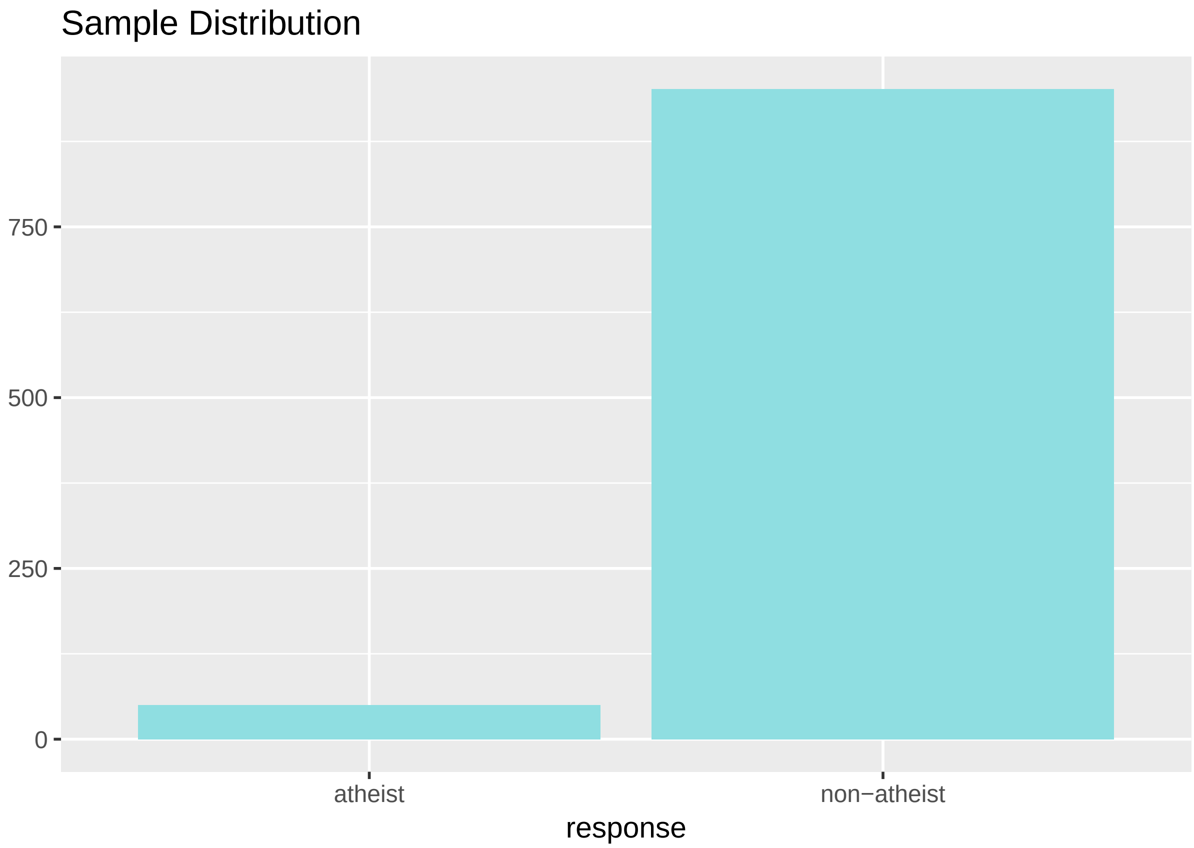
Hence, all conditions are met to construct a 95% confidence interval for the proportion of atheists in United States in 2012.

2. Calculate 95% confidence interval

If the conditions for inference are reasonable, we can either calculate the standard error and construct the interval by hand, or allow the `inference` function to do it for us.

```
inference(y = response, data = us12, statistic = "proportion", type = "ci", method = "theoretical", suc
```

```
## Single categorical variable, success: atheist
## n = 1002, p-hat = 0.0499
## 95% CI: (0.0364 , 0.0634)
```



Note that since the goal is to construct an interval estimate for a proportion, it's necessary to specify what constitutes a "success", which here is a response of `atheist`.

Although formal confidence intervals and hypothesis tests don't show up in the report, suggestions of inference appear at the bottom of page 7: "In general, the error margin for surveys of this kind is ± 3-5% at 95% confidence."

What is a margin of error?

Margin of error (ME) is the error that occurs simply because the researchers are not asking everyone. That is, the surveys are based on information collected from a sample of individuals, not the entire population. Hence, the margin of error measures the maximum amount by which the sample results are expected to differ from those of the actual population.

From the R output above, we can easily calculate the margin of error from confidence interval we obtained to compare it with the one shown in the report.

$$\hat{p} \pm ME \Rightarrow \hat{p} + ME = 0.0634, \quad \hat{p} - ME = 0.0364$$

$$ME = \frac{0.0634 - 0.0364}{2} = 0.0135.$$

```
# Calculate ME from R output
(0.0634 - 0.0364)/2
```

```
## [1] 0.0135
```

Hence, **The margin of error for the estimate of the proportion of atheists in the US in 2012 is 0.0135.**

Similarly, we can use the inference function to calculate confidence intervals for the proportion of atheists in 2012 in two other countries, Peru and Canada, and report the associated margins of error.

Calculate confidence intervals for the proportion of atheists in 2012 in Peru

```
peru12 <- atheism %>%
  filter(nationality == "Peru" , atheism$year == "2012")
```

1. Check conditions for inference

The conditions for inference to construct a 95% confidence interval for the proportion of atheists in the Peru in 2012 are met:

- Since the study uses stratified sampling, and the sample size is 1,207 less than 10% of popuation, it is reasonable to assume that observations are **independent**.
- There are 36 successes (atheist) and 1171 failures (non-atheist). They are both larger than 10 so **the success-failure condition is satisfied.**

```
# Count the numbers in each category
peru12 %>%
     group_by(peru12$response) %>%
     summarize(counts = n())
```

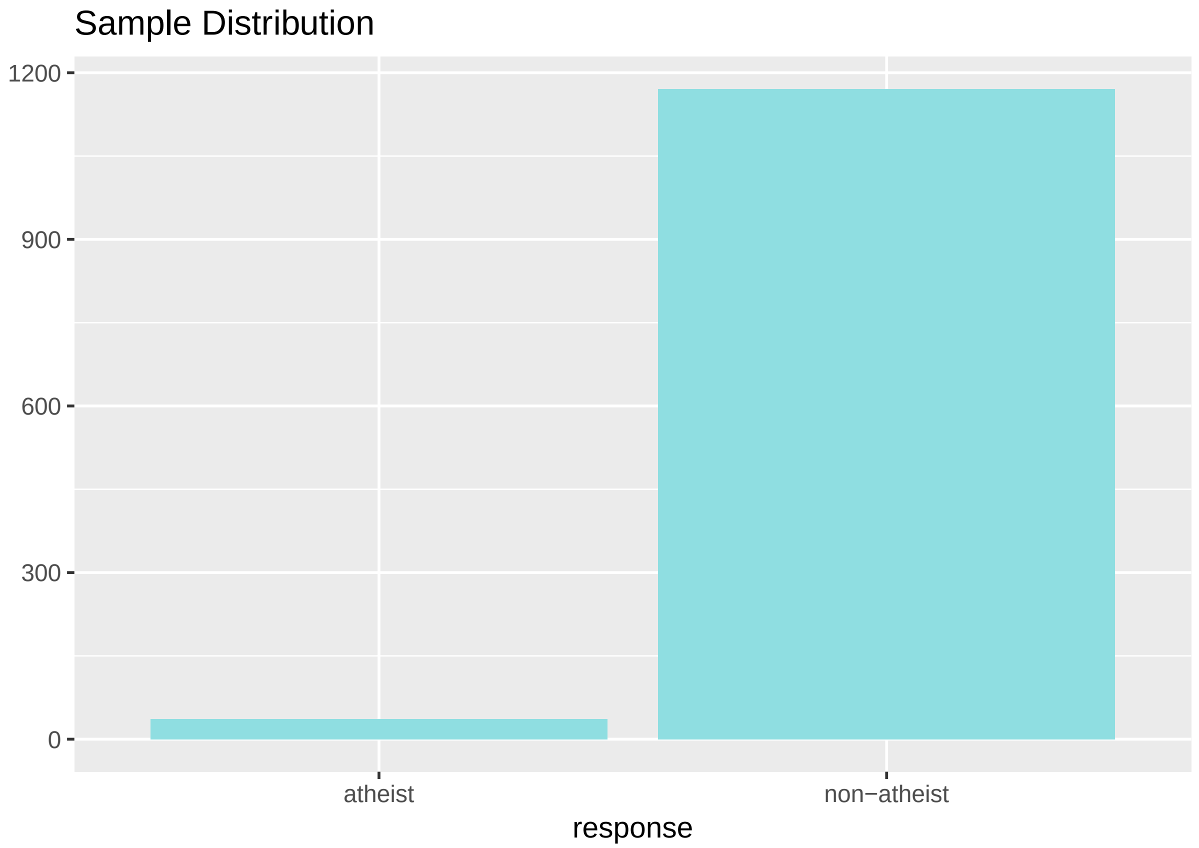
```
## # A tibble: 2 x 2
##   `peru12$response` counts
##   <fct>              <int>
## 1 atheist               36
## 2 non-atheist         1171
```

2. Calculate 95% confidence interval

Similarly, we use `inference` function to calculate confidence interval for the proportion of atheists in 2012 in Peru.

```
inference(y = response, data = peru12, statistic = "proportion", type = "ci", method = "theoretical", su
```

```
## Single categorical variable, success: atheist
## n = 1207, p-hat = 0.0298
## 95% CI: (0.0202 , 0.0394)
```



From the R output above, we can calculate the margin of error from confidence interval we obtained below:

$$ME_{peru} = \frac{0.0394 - 0.0202}{2} = 0.0096.$$

Calculate confidence intervals for the proportion of atheists in 2012 in Canada

```
ca12 <- atheism %>%
  filter(nationality == "Canada" , atheism$year == "2012")
```

1. Check conditions for inference

The conditions for inference to construct a 95% confidence interval for the proportion of atheists in Canada in 2012 are met:

- Since the study uses stratified sampling, and the sample size is 1,002 less than 10% of popuation, it is reasonable to assume that observations are **independent**.
- There are 90 successes (atheist) and 912 failures (non-atheist). They are both larger than 10 so **the success-failure condition is satisfied.**

```
# Count the numbers in each category
ca12 %>%
     group_by(ca12$response) %>%
     summarize(counts = n())
```

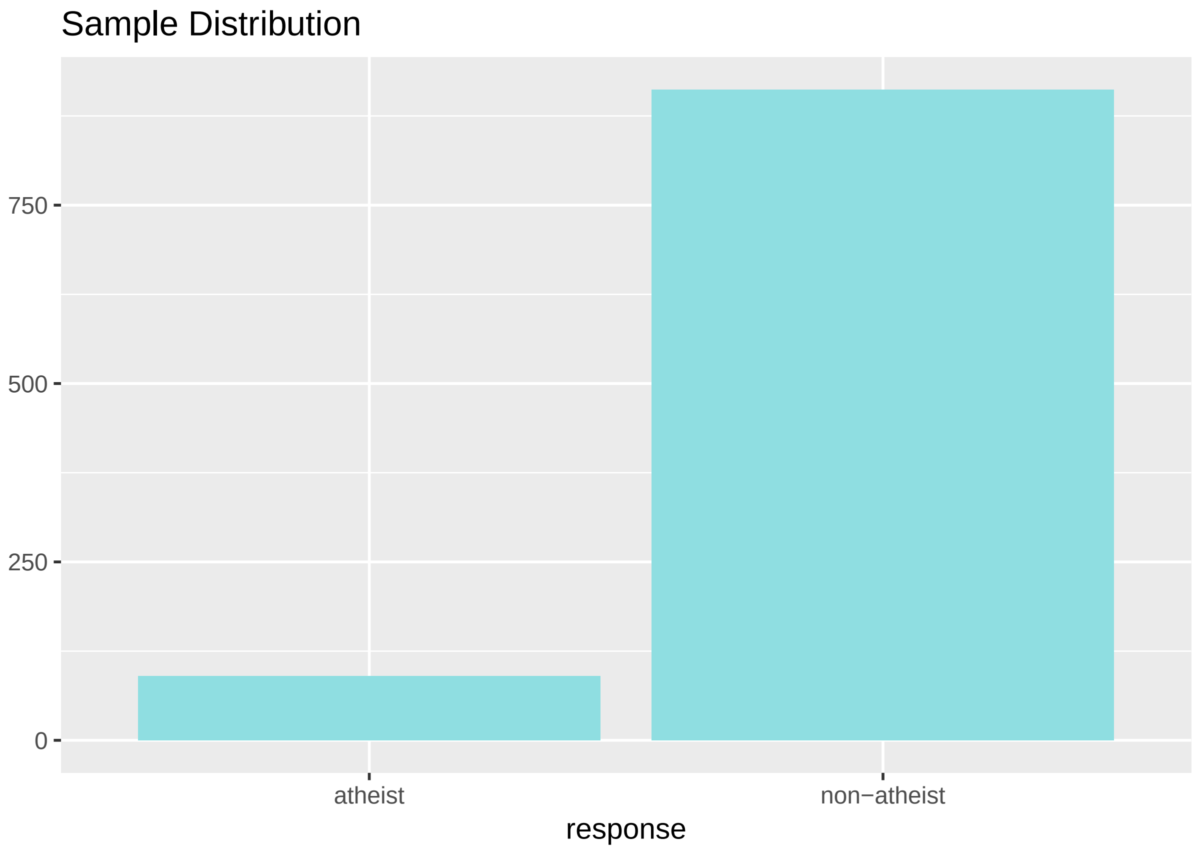
```
## # A tibble: 2 x 2
##   `ca12$response` counts
##   <fct>            <int>
## 1 atheist             90
## 2 non-atheist        912
```

2. Calculate 95% confidence interval

We use `inference` function to calculate confidence interval for the proportion of atheists in 2012 in Canada.

```
inference(y = response, data = ca12, statistic = "proportion", type = "ci", method = "theoretical", suc
```

```
## Single categorical variable, success: atheist
## n = 1002, p-hat = 0.0898
## 95% CI: (0.0721 , 0.1075)
```



From the R output above, we can calculate the margin of error from confidence interval we obtained below:

$$ME_{ca} = \frac{0.1075 - 0.0721}{2} = 0.0177.$$

How does the proportion affect the margin of error?

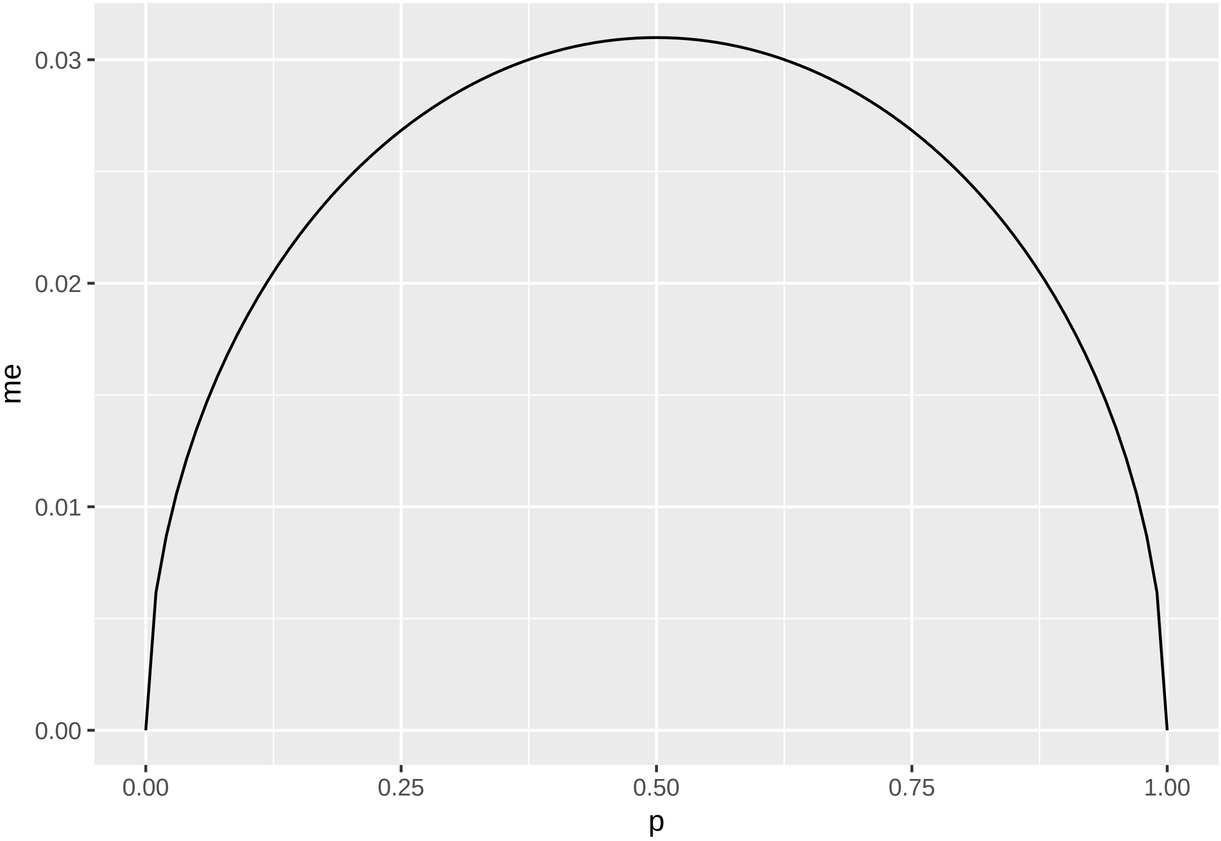
Imagine you've set out to survey 1000 people on two questions: are you female? and are you left-handed? Since both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! **While the margin of error does change with sample size, it is also**

affected by the proportion. As shown in previous section, the sample sizes in United States and Canada in 2012 are identical in the dataset but we still obtained different margin of error.

Think back to the formula for the standard error: $SE = \sqrt{p(1-p)/n}$. This is then used in the formula for the margin of error for a 95% confidence interval: $ME = 1.96 \times SE = 1.96 \times \sqrt{p(1-p)/n}$. Since the population proportion p is in this ME formula, it should make sense that the margin of error is in some way dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p.

The first step is to make a vector `p` that is a sequence from 0 to 1 with each number separated by 0.01. We can then create a vector of the margin of error (`me`) associated with each of these values of `p` using the familiar approximate formula ($ME = 1.96 \times SE$). Lastly, we plot the two vectors against each other to reveal their relationship.

```
d <- data.frame(p <- seq(0, 1, 0.01))
n <- 1000
d <- d %>%
  mutate(me = 1.96*sqrt(p*(1 - p)/n))
ggplot(d, aes(x = p, y = me)) +
  geom_line()
```



From above figure, we can observe the relationship between p and ME in three extreme cases: the ME reaches a minimum at $p = 0$, the ME reaches a minimum at $p = 1$, and the ME is maximized when $p = 0.5$.

From above figure, we can see that when $p = 0.5$, margin of error (ME) has a maximum with given sample size n and 95% confidence. We can also prove this observation from calculus in the following:

$$ME = z^* \times SE$$

where $z^* = 1.96$ for 95% confidence and SE is given by $SE = \sqrt{p(1-p)/n}$. Now we fix the sample size n and want to maximize ME with respect to p so we calculate the first derivative of ME with respect to p equal to 0.

$$\frac{d(ME)}{dp} = 0 \Rightarrow \frac{d(ME)}{dp} = \frac{z^*}{\sqrt{n}}\frac{d}{dp}[p(1-p)]^{1/2} = \frac{z^*}{\sqrt{n}} \times \frac{1}{2}\frac{[(1-p)+p(-1)]}{\sqrt{p(1-p)}} = 0$$

Since $\sqrt{p(1-p)} \geq 0$, the solution for p in above equation is $p = \frac{1}{2}$. That means that ME has extremum at $p = 0.5$. Because ME is a concave function, at $p = 0.5$ ME has maximum.

Are there any changes in atheism index between 2005 and 2012?

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005. We assume here that sample sizes have remained the same. Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

Answer the following two questions using the `inference` function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012?

1. Set the hypothesis

- H_0: Spain has not seen a change in its atheism index between 2005 and 2012.

$$p_{2005} - p_{2012} = 0.$$

- H_A: Spain has seen a change in its atheism index between 2005 and 2012.

$$p_{2005} - p_{2012} \neq 0.$$

  We will carry out two-sided hypothesis testing for comparing two independent proportions.

2. Check conditions for inference

```
sp05_12 <- atheism %>%
  filter(nationality == "Spain" , (atheism$year == "2012") | (atheism$year == "2005") )

# Convert year from integer to factor
sp05_12$year <- as.factor(sp05_12$year)

# Count number in each category in each year
sp05_12 %>%
     group_by(years = sp05_12$year, response = sp05_12$response) %>%
     summarize(counts = n())
```

```
## # A tibble: 4 x 3
## # Groups:   years [?]
##   years response    counts
##   <fct> <fct>        <int>
## 1 2005  atheist        115
## 2 2005  non-atheist   1031
## 3 2012  atheist        103
## 4 2012  non-atheist   1042
```

- Since the study uses stratified sampling, and the sample sizes in 2005 and 2012 are 1,146 and 1,145, respectively, less than 10% of popuation, it is reasonable to assume that observations are **independent within groups and between groups**.

- Since when we do hypothesis testing, we assume null hypothesis is true ($p_{2005} = p_{2012}$), we use pooled proportion to examine success-failure condition. As shown in the following, **the success-failure condition is satisfied in each group.**

$$p_{pool} = \frac{115 + 103}{1146 + 1145} = 0.095$$

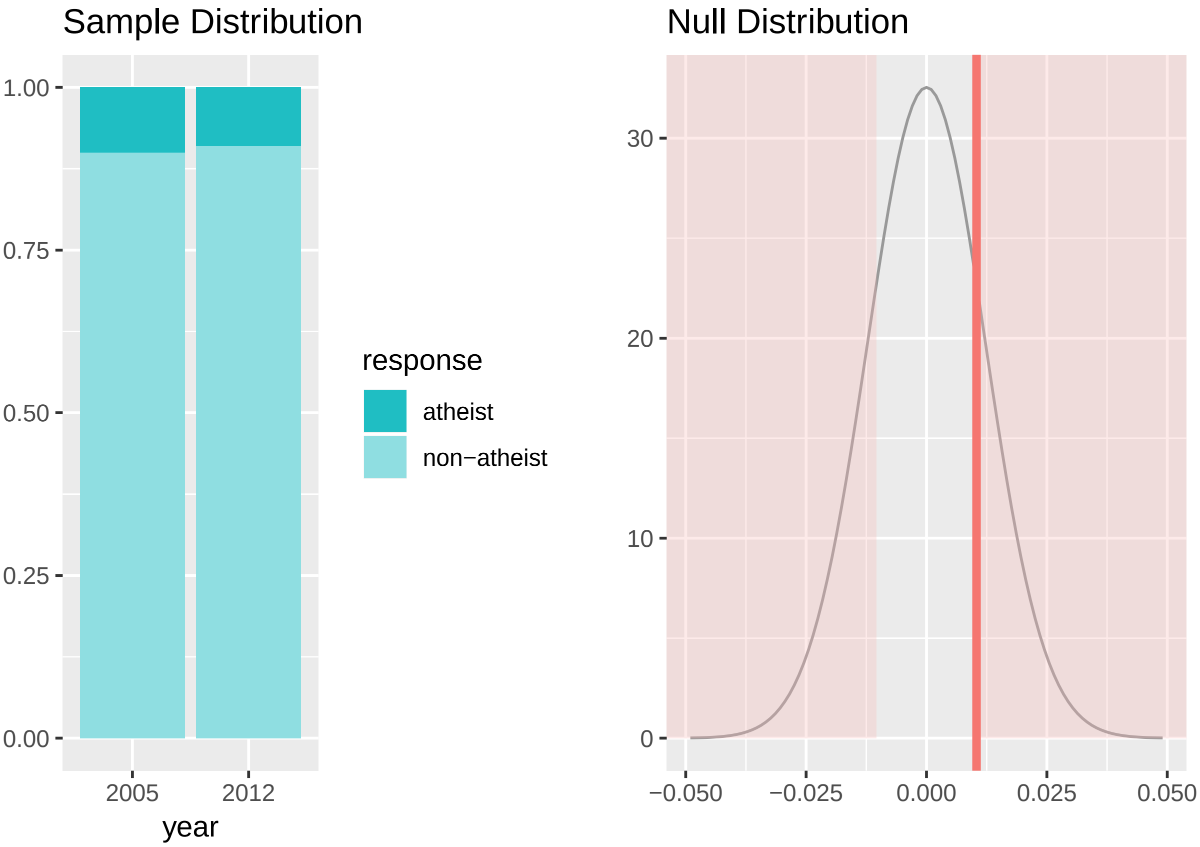
$$1146 \times 0.095 \geq 10, \quad 1146 \times (1 - 0.095) \geq 10$$

$$1145 \times 0.095 \geq 10, \quad 1145 \times (1 - 0.095) \geq 10$$

3. Calculate p-value

```
inference(y = response, x = year, data = sp05_12, statistic = "proportion", type = "ht", method = "theo
```

```
## Response variable: categorical (2 levels, success: atheist)
## Explanatory variable: categorical (2 levels)
## n_2005 = 1146, p_hat_2005 = 0.1003
## n_2012 = 1145, p_hat_2012 = 0.09
## H0: p_2005 =  p_2012
## HA: p_2005 != p_2012
## z = 0.8476
## p_value = 0.3966
```



Since p-value is 0.3966 greater than significance level $\alpha = 0.05$, we fail to reject the null hypothesis. Hence, **there is no convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012.**

Is there convincing evidence that United States has seen a change in its atheism index between 2005 and 2012?

1. Set the hypothesis

- H_0: United States has not seen a change in its atheism index between 2005 and 2012.

$$p_{2005} - p_{2012} = 0.$$

- H_A: United States has seen a change in its atheism index between 2005 and 2012.

$$p_{2005} - p_{2012} \neq 0.$$

  We will carry out two-sided hypothesis testing for comparing two independent proportions.

2. Check conditions for inference

```
us05_12 <- atheism %>%
  filter(nationality == "United States" , (atheism$year == "2012") | (atheism$year == "2005") )

# Convert year from integer to factor
us05_12$year <- as.factor(us05_12$year)

# Count number in each category in each year
us05_12 %>%
     group_by(years = us05_12$year, response = us05_12$response) %>%
     summarize(counts = n())
```

```
## # A tibble: 4 x 3
## # Groups:   years [?]
##   years response    counts
##   <fct> <fct>        <int>
## 1 2005  atheist         10
## 2 2005  non-atheist    992
## 3 2012  atheist         50
## 4 2012  non-atheist    952
```

- Since the study uses stratified sampling, and the sample sizes in 2005 and 2012 are 1,002 and 1,002, respectively, less than 10% of popuation, it is reasonable to assume that observations are **independent within groups and between groups**.
- Since when we do hypothesis testing, we assume null hypothesis is true ($p_{2005} = p_{2012}$), we use pooled proportion to examine success-failure condition. As shown in the following, **the success-failure condition is satisfied in each group.**

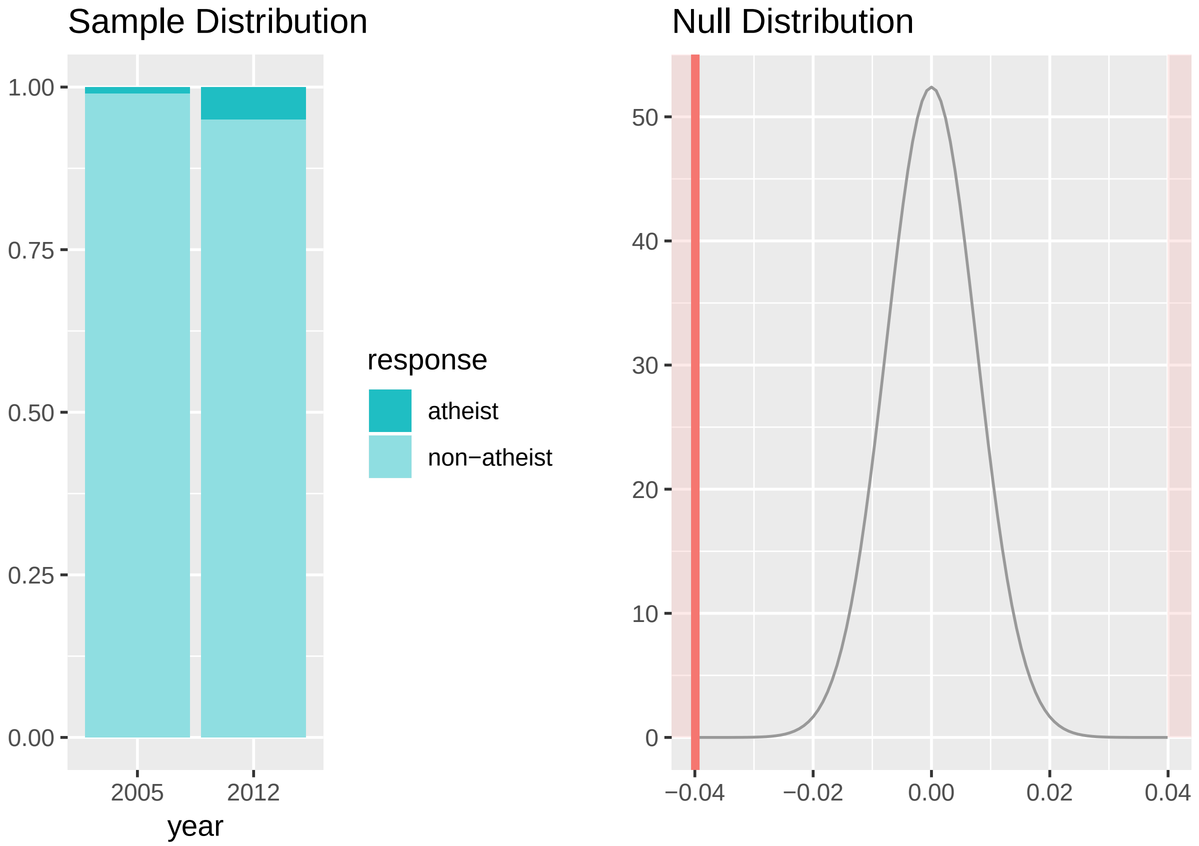
$$p_{pool} = \frac{10 + 50}{1002 + 1002} = 0.02994$$

$$1002 \times 0.02994 \geq 10, \quad 1002 \times (1 - 0.02994) \geq 10$$

$$1002 \times 0.02994 \geq 10, \quad 1002 \times (1 - 0.02994) \geq 10$$

3. Calculate p-value

```
inference(y = response, x = year, data = us05_12, statistic = "proportion", type = "ht", method = "theo
## Response variable: categorical (2 levels, success: atheist)
## Explanatory variable: categorical (2 levels)
## n_2005 = 1002, p_hat_2005 = 0.01
## n_2012 = 1002, p_hat_2012 = 0.0499
## H0: p_2005 =  p_2012
## HA: p_2005 != p_2012
## z = -5.2431
## p_value = < 0.0001
```



Since p-value is less than significance level $\alpha = 0.05$, we reject the null hypothesis and accept the alternative hypothesis. Hence, **there is convincing evidence that United States has seen a change in its atheism index between 2005 and 2012.**

If in fact there has been no change in the atheism index in the countries listed in Table 4, how many of those countries would you expect to detect a change (at a significance level of 0.05) simply by chance?

Recall that the **Type 1 error** is the probability of rejecting null hypothesis, H_0, when H_0 is true,

$$\text{Type 1 error} = P(\text{reject } H_0 | H_0 \text{ is true}).$$

Actually, Type 1 error is the significance level α we set in hypothesis testing. Here, we set $\alpha = 0.05$ or 5%. Hence, if in fact there has been no change in the atheism index in the 39 countries listed in Table 4 (H_0 is true.), then there would be $0.05 \times 39 = 1.95$ countries we expect to detect a change simply by chance.

How to determine the least sample size we need to carry out the survey such that the margin of error is no greater than 1% with 95% confidence?

Suppose you're hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. According to the guidelines, the estimate must have a margin of error no greater than 1% with 95% confidence. You have no idea what to expect for p. How many people would you have to sample to ensure that you are within the guidelines?

Recall that the margin of error (ME) is given by the formula,

$$ME = z^* \times SE = z^* \times \sqrt{\frac{p(1-p)}{n}}.$$

For 95% confidence, $z^* = 1.96$. Hence, the ME depends on p and n. As we have discussed before, when $p = 0.5$ with given n, ME has maximum. Since we don't want to underestimate n, we adopt $p = 0.5$ to estimate the least sample size n we need such that ME is no greater than 0.01.

$$ME \leq 0.01 \Rightarrow z^* \sqrt{\frac{p(1-p)}{n}} \leq 0.01 \Rightarrow n \geq \left[\frac{z^*}{0.01}\right]^2 p(1-p)$$

where $z^* = 1.96$ for 95% confidence and $p = 0.5$. Thus, the least n is 9604.

```
# Calculae the least sample size n
(1.96*0.5/0.01)**2
```

```
## [1] 9604
```

This was written by Hsuan-Hao Fan.